

# Universal Bayesian Measures and Universal Histogram Sequences\*

Joe Suzuki <sup>†</sup>

## Abstract

Consider universal data compression: the length  $l(x^n)$  of sequence  $x^n \in A^n$  with finite alphabet  $A$  and length  $n$  satisfies Kraft's inequality over  $A^n$ , and  $-\frac{1}{n} \log \frac{P^n(x^n)}{Q^n(x^n)}$  almost surely converges to zero as  $n$  grows for the  $Q^n(x^n) = 2^{-l(x^n)}$  and any stationary ergodic source  $P$ . In this paper, we say such a  $Q$  is a universal Bayesian measure. We generalize the notion to the sources in which the random variables may be either discrete, continuous, or none of them. The basic idea is due to Boris Ryabko who utilized model weighting over histograms that approximate  $P$ , assuming that a density function of  $P$  exists. However, the range of  $P$  depends on the choice of the histogram sequence. The universal Bayesian measure constructed in this paper overcomes the drawbacks and has many applications to infer relation among random variables, and extends the application area of the minimum description length principle.

keywords: universal coding, Radon-Nikodym, Bayesian measure, estimation, density function

## 1 Introduction

Suppose we wish to know if discrete random variables  $X, Y$  are independent ( $X \perp\!\!\!\perp Y$ ) given  $n$  pairs of examples  $\{(x_i, y_i)\}_{i=1}^n$  emitted by  $(X, Y)$ . If the probabilities of  $x^n = (x_1, \dots, x_n)$ ,  $y^n = (y_1, \dots, y_n)$ , and  $(x^n, y^n)$  are expressed by  $P_X^n(x^n|\theta_X)$ ,  $P_Y^n(y^n|\theta_Y)$ , and  $P_{XY}^n(x^n, y^n|\theta_{XY})$ , respectively, using unknown parameters  $\theta_X, \theta_Y, \theta_{XY}$ , one way to deal with this problem is to decide  $X \perp\!\!\!\perp Y$  if and only if

$$pQ_X^n(x^n)Q_Y^n(y^n) \geq (1-p)Q_{XY}^n(x^n, y^n) ,$$

where  $p$  is the prior probability of  $X \perp\!\!\!\perp Y$ , and the three values are defined by

$$\begin{aligned} Q_X^n(x^n) &:= \int P^n(x^n|\theta_X)w_X(\theta_X)d\theta_X , \\ Q_Y^n(y^n) &:= \int P^n(y^n|\theta_Y)w_Y(\theta_Y)d\theta_Y , \\ Q_{XY}^n(x^n, y^n) &:= \int P^n(x^n, y^n|\theta_{XY})w_{XY}(\theta_{XY})d\theta_{XY} \end{aligned} \tag{1}$$

using weights  $w_X, w_Y, w_{XY}$  over the parameters  $\theta_X, \theta_Y, \theta_{XY}$ , respectively.

To this end, let  $A$  be the finite set in which  $X$  takes values. There are many options of  $Q_X$  such that

$$\sum_{x^n \in A^n} Q_X^n(x^n) \leq 1 . \tag{2}$$

For example<sup>1</sup>,  $Q_X^n(x^n) = |A|^{-n}$  for  $x^n \in A^n$  satisfies the condition. However, such a  $Q_X$  cannot be an alternative of  $P$  for large  $n$  because  $Q_X^n$  does not converges to  $P^n$  in any sense. On the other

\*This paper was partially presented at IEEE International Symposium on Information Theory, Istanbul, Turkey, July, 2013.

<sup>†</sup>J. Suzuki is with the Department of Mathematics, Osaka University, Toyonaka, Osaka, 560-0043, JAPAN, e-mail: suzuki@math.sci.osaka-u.ac.jp

<sup>1</sup> $|A|$  denotes the cardinality of set  $A$ .

hand, if we choose  $w_X(\theta_X) \propto \prod_{x \in A} \theta_x^{-a[x]}$  with constants  $(a[x] = \frac{1}{2})_{x \in A}$  (Krichevsky-Trofimov [3]), then the quantity  $-\frac{1}{n} \log Q_X^n(x^n)$  almost surely converges to its entropy  $H(\theta_X)$  for any independent and identically distributed (i.i.d) source  $P^n(x^n|\theta_X) = \prod_{x \in A} \theta_x^{-c[x]}$  with parameters  $\theta = (\theta_x)_{x \in A}$  and frequencies  $(c[x])_{x \in A}$  in  $x^n \in A^n$  [6]. Furthermore, the Shannon-McMillian-Breiman theorem [2] states that  $-\frac{1}{n} \log P^n(x^n|\theta_X)$  almost surely converges to  $H(\theta_X)$  for any stationary ergodic source  $\theta_X$ , so that almost surely

$$\frac{1}{n} \log \frac{P_X^n(x^n)}{Q_X^n(x^n)} \rightarrow 0 \quad (3)$$

if we write  $P^n(x^n|\theta_X)$  by  $P_X^n(x^n)$ . In this paper, we say such a  $Q_X$  satisfying (2)(3) to be a *universal Bayesian measure* associated with finite set  $A$ . From the above discussion, we can say that a universal Bayesian measure exists for finite sources.

However, what if  $X, Y$  are arbitrary without assuming they are discrete? Recently, for random variable  $X$  such that its density function  $f_X$  exists, Boris Ryabko [5] proved that there exists  $g_X$  such that

$$\int_{x^n \in \mathbb{R}^n} g_X^n(x^n) \leq 1$$

and

$$\frac{1}{n} \log \frac{f_X^n(x^n)}{g_X^n(x^n)} \rightarrow 0. \quad (4)$$

for any  $f_X$  satisfying a condition which will be specified in the later sections.

In addition, in order to decide whether  $X \perp\!\!\!\perp Y$  or not is made, we need to construct Bayesian measures  $Q_{XY}$  and  $g_{XY}$  for two variables  $X, Y$  extending  $Q_X$  and  $g_X$  for one variable  $X$ .

We admire Ryabko's original work [5], and admit that the basic idea was already there. However, we need to seek further generalizations for practical development of the theory. The purposes of this paper are

1. to remove the constraint that  $X$  should be either discrete or continuous to obtain a general form of universality containing (3)(4) as special cases;
2. to remove the condition that Ryabko [5] posed; and
3. to construct universal measures for more than one variables,

so that we establish that a universal Bayesian measure unconditionally exists for any stationary ergodic random variable which may be either discrete, continuous, or none of them. Once we can deal with universal Bayesian measures for more than one random variables, we can infer relation among them from given examples.

For simplicity, in this paper, we assume that the underlying source is i.i.d. although the discussion will hold for stationary ergodic sources.

This paper is organized as follows: Section 2 gives a basic material and background of this paper. In Sections 3,4,5, we solve the three problems above in the form of Theorems 1,2,3, respectively. Section 6 concludes this paper by suggesting applications to show how significant the three results are. Throughout the paper, we denote the entire real, rational, integer, and natural numbers by  $\mathbb{R}, \mathbb{Q}, \mathbb{Z}$ , and  $\mathbb{N}$ , respectively.

## 2 Preliminaries

### 2.1 Ryabko's measure

Let  $X$  be a random variable for which a density function  $f_X$  exists, and  $A$  the set in which  $X$  takes values. Let  $\{A_j\}_{j=0}^{\infty}$  be such that  $A_0 := \{A\}$  and that  $A_{j+1}$  is a refinement of  $A_j$ .

**Example 1** If<sup>2</sup>  $A = [0, 1)$ , the sequence  $A_0 = \{[0, 1)\}$

$$A_1 = \{[0, 1/2), [1/2, 1)\}$$

$$A_2 = \{[0, 1/4), [1/4, 1/2), [1/2, 3/4), [3/4, 1)\}$$

...

$$A_j = \{[0, 2^{-(j-1)}), [2^{-(j-1)}, 2 \cdot 2^{-(j-1)}),$$

$$\dots, [(2^{j-1} - 1)2^{-(j-1)}, 1)\}$$

...

satisfies the condition.

Let  $\lambda$  be the Lebesgue measure. For example, if  $a = [b, c)$ , then  $\lambda(a) = c - b$ . For each  $j = 1, 2, \dots$ , let  $s_j : A \rightarrow A_j$  be such that  $x \in a \in A_j \implies s_j(x) = a$ ,

$$P_j(a) := \int_{x \in a} f_X(x) dx, \quad a \in A_j$$

and

$$f_j(x) := \frac{P_j(s_j(x))}{\lambda(s_j(x))}, \quad x \in A.$$

Then, we consult the following lemma:

**Lemma 1 ([6])** For any finite set  $A$ , a universal Bayesian measure associated with  $A$ .

(For proof, see Appendix A.)

Since each  $A_j$  is a finite set, we can construct a universal Bayesian measure  $Q_j$  associated with  $A_j$ .

Suppose we are given  $x^n = (x_1, \dots, x_n) \in A^n$  such that  $(s_j(x_1), \dots, s_j(x_n)) = (a_1, \dots, a_n) \in A_j^n$ . Then, for each  $j = 1, 2, \dots$ , we approximate the value of the approximated density function

$$f_j^n(x^n) := f_j(x_1) \cdots f_j(x_n) = \frac{P_j(a_1) \cdots P_j(a_n)}{\lambda(a_1) \cdots \lambda(a_n)}$$

by

$$g_j^n(x^n) := \frac{Q_j^n(a_1, \dots, a_n)}{\lambda(a_1) \cdots \lambda(a_n)}.$$

Let  $\{\omega_j\}_{j=1}^\infty$  be such that  $\sum \omega_j = 1$ ,  $\omega_j > 0$ . Ryabko proved [5] that  $g_X^n(x^n) = \sum_{j=0}^\infty \omega_j g_j^n(x^n)$  and  $f_X^n(x^n) = f_X(x_1) \cdots f_X(x_n)$  satisfy (4) for any  $f$  such that  $D(f_X || f_j) \rightarrow 0$  as  $j \rightarrow \infty$ , where  $D(f || g)$  is the Kullback-Leibler divergence of  $f$  from  $g$ :

$$D(f || g) := \int_{x \in A} f(x) \log \frac{f(x)}{g(x)} dx.$$

We should notice that the set  $\{f_X | D(f_X || f_j) \rightarrow 0 \text{ as } j \rightarrow \infty\}$  depends on the histogram sequence  $\{A_j\}$ . In this sense, Ryabko's measure is a universal Bayesian measure w.r.t. a specific  $\{A_j\}$ . In this paper, we refer this constraint to Ryabko's condition.

## 2.2 Exactly when a density function exists ?

Let  $\mathcal{B}$  the entire Borel sets of  $\mathbb{R}$ . Formally,  $X$  is a random variable if  $(X \in D) \in \mathcal{F}$  for any  $D \in \mathcal{B}$  for the underlying probability space  $(\Omega, \mathcal{F}, P)$ . Given random variable  $X$ , a necessary and sufficient condition is available that its density function exists:

**Example 2** The following two condition are equivalent [1]:

1. For any  $D \in \mathcal{B}$ , there exists  $f_X : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that

$$\mu_X(D) := P(X \in D) = \int_{x \in D} f_X(x) dx.$$

---

<sup>2</sup>For  $b < c$ ,  $[b, c)$  denotes the set  $\{x \in \mathbb{R} | b \leq x < c\}$ .

2. For any  $D \in \mathcal{B}$ ,

$$\lambda(D) := \int_{x \in D} dx = 0 \implies \mu_X(D) = 0$$

Then, such an  $f_X$  (density function) is obtained by  $f_X(x) = \frac{dF_X(x)}{dx}$ , where  $F_X$  is the distribution function of  $X$ .

**Example 3** We can check the following two conditions are equivalent:

1. For any  $D \in \mathcal{B}$ , there exists  $f_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that

$$\mu_Y(D) := P(Y \in D) = \sum_{y \in D \cap \mathbb{Z}} f_Y(y)$$

2. For any  $D \in \mathcal{B}$ ,

$$\eta(D) := |D \cap \mathbb{Z}| = 0 \implies \mu_Y(D) = 0 .$$

Then, such an  $f_Y$  is obtained by  $f_Y(y) = P(Y = y) = \mu_Y(\{y\})$  for  $y \in \mathbb{Z}$  ( $f_Y(y)$  may take any value for  $y \notin \mathbb{Z}$ ).

Notice that a density function in a generalized sense exists even if the random variable is discrete.

Let  $\eta$  be a  $\sigma$ -finite measure, i.e. there exists  $\{A_j\}$  such that  $A_j \in \mathcal{F}$ ,  $\cup_j A_j = \Omega$  and  $\eta(A_i) < \infty$  for measure space  $(\Omega, \mathcal{F})$ . For example,  $\lambda$  and  $\eta$  in Examples 2 and 3 are both  $\sigma$ -finite because  $A_j = [j, j+1)$ ,  $A_j \in \mathcal{B}$ ,  $\cup_j A_j = \mathbb{R}$  and  $\lambda(A_j) = \eta(A_j) = 1$ .

For any random variable  $X$  which is either discrete or continuous or none of them, there exists a density function  $f_Z$  w.r.t.  $\eta$  as long as  $\mu_Z$  is absolutely continuous w.r.t.  $\eta$ , where  $\mu_Z(D) := P(Z \in D)$  for  $D \in \mathcal{B}$ :

**Lemma 2 (Radon-Nikodym [1])** Let  $\mu, \eta$  be  $\sigma$ -finite measures for measure space  $(\Omega, \mathcal{F})$ . Then, the following two conditions are equivalent:

1. For any  $A \in \mathcal{F}$ , there exists nonnegative  $f$  such that  $\mu(A) = \int_A f(t) d\eta(t)$ .
2. For any  $A \in \mathcal{F}$ ,  $\eta(A) = 0 \implies \mu(A) = 0$

If the condition in Lemma 1 is met, we say that  $\mu$  is *absolutely continuous* w.r.t.  $\eta$  and write  $\mu \ll \eta$ .

We notice that the integral in the lemma is the Lebesgue integral that takes the value

$$\sup_i \sum [\inf_{\omega \in A_i} f(\omega)] \eta(A_i)$$

for  $\mathcal{F}$ -measurable function  $f$ , i.e.,  $\{\omega \in \Omega | f(\omega) \in D\} \in \mathcal{F}$  for any  $D \in \mathcal{B}$ , where the supreme is over  $\{A_i\}$  such that  $A_i \cap A_j = \emptyset$  for  $i \neq j$  and  $\cup A_i = \Omega$ , and contains the Riemann integrals and summations as in Examples 2 and 3 as special cases. Such a density function  $f$  in the generalized sense is called a *Radon-Nikodym derivative*.

**Example 4** Let  $\lambda$  and  $\eta$  be as in Examples 2 and 3, respectively, and  $\xi(D) := \lambda(D) + \eta(D)$  for  $D \in \mathcal{B}$ . Then, the following two conditions are equivalent:

1. For any  $D \in \mathcal{B}$ , there exists  $f_Z : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that

$$\begin{aligned} \mu_Z(D) &:= P(Z \in D) \\ &= \int_{Z \in D} f_Z(z) dz + \sum_{z \in D \cap \mathbb{Z}} f_Z(z) \\ &= \int_{z \in D} f_Z(z) d\xi(z) \end{aligned}$$

2. For any  $D \in \mathcal{B}$ ,

$$\xi(D) = 0 \implies \mu_Z(D) = 0 .$$

Notice that a density function in the generalized sense exists even when the random variable is not either discrete or continuous.

### 3 Estimating a density function in the generalized sense

Based on the discussion in Section 2.2, we generalize the result in Section 2.1 to the one that does not assume the random variable to be either discrete or continuous.

Let  $\eta$  be a  $\sigma$ -finite measure. Let  $Y$  be a random variable such that  $\mu_Y \ll \eta$  for  $\mu_Y(D) = P(Y \in D)$ ,  $D \in \mathcal{B}$ , and  $B$  the set in which  $X$  takes values. Let  $\{B_k\}_{k=0}^\infty$  be such that  $B_0 := \{B\}$  and that  $B_{k+1}$  is a refinement of  $B_k$ .

**Example 5** Let  $\eta(\{h\}) := \frac{1}{h(h+1)}$  for  $h \in B = \mathbb{N} = \{1, 2, \dots\}$ . We assume that  $\mu(\{h\}) > 0$  only if  $h \in B$ . Then,  $\mu \ll \eta$  for  $\mu_Y(D) = P(Y \in D)$ ,  $D \in \mathcal{B}$ , and from Lemma 2, there exists  $f_Y$  such that

$$\mu_Y(D) = \sum_{h \in D} f_Y(h) \eta(\{h\}) .$$

In fact,

$$f_Y(h) = \frac{\mu_Y(\{h\})}{\eta(\{h\})} = h(h+1)\mu_Y(\{h\})$$

satisfies the property. For  $\{B_k\}$ , the following sequence satisfies the condition:

$$\begin{aligned} B_1 &:= \{\{1\}, \{2, 3, \dots\}\} \\ B_2 &:= \{\{1\}, \{2\}, \{3, 4, \dots\}\} \\ &\dots \\ B_k &:= \{\{1\}, \{2\}, \dots, \{k\}, \{k+1, k+2, \dots\}\} \\ &\dots \end{aligned}$$

For each  $k = 1, 2, \dots$ , let  $t_k : B \rightarrow B_k$  be such that  $y \in b \in B_k \implies t_k(y) = b$ ,

$$P_k(b) := \int_{y \in b} f_Y(y) d\eta(y)$$

for  $b \in B_k$ , and

$$f_k(y) := \frac{P_k(t_k(y))}{\eta(t_k(y))}$$

for  $y \in B$ . Since  $B_k$  is a finite set, we can construct a universal Bayesian measure  $Q_k$  associated with  $B_k$ .

Suppose we are given  $y^n = (y_1, \dots, y_n) \in B^n$  such that  $(t_k(y_1), \dots, t_k(y_n)) = (b_1, \dots, b_n) \in B_k^n$  for  $k = 1, 2, \dots$ . Then, for each  $k = 1, 2, \dots$ , we estimate

$$f_k^n(y^n) := f_k(y_1) \cdots f_k(y_n) = \frac{P_k(b_1) \cdots P_k(b_n)}{\eta(b_1) \cdots \eta(b_n)} .$$

by

$$g_k^n(y^n) := \frac{Q_k^n(b_1, \dots, b_n)}{\eta(b_1) \cdots \eta(b_n)} .$$

Let  $\{\omega_k\}_{k=1}^\infty$  be such that  $\sum \omega_k = 1$ ,  $\omega_k > 0$ . We claim that  $g_X^n(y^n) = \sum_{k=0}^\infty \omega_k g_k^n(y^n)$  and  $f_Y^n(y^n) = f_Y(y_1) \cdots f_Y(y_n)$  satisfies (4) for any  $f$  such that  $D(f_Y || f_k) \rightarrow 0$  as  $k \rightarrow \infty$ , where  $D(f || g)$  is the Kullback-Leibler divergence of  $f$  from  $g$ :

$$D(f || g) := \int_{y \in B} f(y) \log \frac{f(y)}{g(y)} d\eta(y) .$$

For  $D^n = (D_1, \dots, D_n)$  with  $D_i \in \mathcal{B}$ , if we define

$$\mu_Y^n(D^n) := \int_{D^n} f_Y^n(y^n) d\eta^n(y^n) ,$$

$$\nu_Y^n(D^n) := \int_{D^n} g_Y^n(y^n) d\eta^n(y^n) ,$$

and  $\eta^n(D^n) = \prod_{i=1}^n \eta(D_i)$ , then we have

$$\frac{f^n(y^n)}{g^n(y^n)} = \frac{d\mu^n}{d\eta^n}(y^n) / \frac{d\nu^n}{d\eta^n}(y^n) = \frac{d\mu^n}{d\nu^n}(y^n) .$$

The Kullback-Leibler divergence of  $f_k$  from  $f_Y$  becomes

$$\begin{aligned} D(f_Y || f_k) &:= \int f_Y(y) \log \frac{f_Y(y)}{f_k(y)} d\eta(y) \\ &= \int \frac{d\mu_Y}{d\eta}(y) \log \left\{ \frac{d\mu_Y}{d\eta}(y) / \frac{d\mu_k}{d\eta}(y) \right\} d\eta(y) \\ &= \int d\mu_Y(y) \log \frac{d\mu_Y}{d\mu_k}(y) = D(\mu_Y || \mu_k) \end{aligned}$$

We arbitrarily fix  $\{B_k\}_{k=0}^\infty$  so that  $B_0 := \{B\}$  and that  $B_{k+1}$  is a refinement of  $B_k$ . Then, the claim is stated in the following form:

**Theorem 1** If  $\mu_Y \ll \eta$ , there exists a  $\nu_Y^n \ll \eta^n$  such that  $\nu_Y^n(B^n) \leq 1$  and with probability one as  $n \rightarrow \infty$

$$\frac{1}{n} \log \frac{d\mu_Y^n}{d\nu_Y^n}(y^n) \rightarrow 0 \quad (5)$$

for any  $\mu_Y$  such that  $D(\mu_Y || \mu_k) \rightarrow 0$  as  $k \rightarrow \infty$ .

Proof: First, we notice for each  $y^{n-1} = (y_1, \dots, y_{n-1})$

$$\begin{aligned} &\int_{y_n \in B} g_k^n(y^n) d\eta(y_n) \\ &= g_k^{n-1}(y^{n-1}) \int_{y_n \in B} \frac{Q_k^n(y^n | y^{n-1})}{\eta(t_k(y_n))} d\eta(y_n) \\ &= g_k^{n-1}(y^{n-1}) \sum_{b \in B_k} Q_k^n(b | t_k(y_1), \dots, t_k(y_{n-1})) \\ &\leq g_k^{n-1}(y^{n-1}) \end{aligned}$$

Thus,

$$\begin{aligned} \int_{y_n} g_Y^n(y^n) d\eta(y_n) &= \sum_k w_k \int_{y_n} g_k^n(y^n) d\eta(y_n) \\ &\leq \sum_k w_k g_k^{n-1}(y^{n-1}) = g_Y^{n-1}(y^{n-1}) , \end{aligned}$$

so that  $\{Z_n\}$  with  $z_n := \frac{g_Y^n(y^n)}{f_Y^n(y^n)}$  is a super-martingale:

$$\begin{aligned} &E[Z | y^{n-1}] \\ &= \frac{g_Y^{n-1}(y^{n-1})}{f_Y^{n-1}(y^{n-1})} \cdot E\left[ \frac{g_Y^n(y^n)}{g_Y^{n-1}(y^{n-1})} \cdot \frac{1}{f_Y(y_n)} \right] \\ &\leq z_{n-1} \cdot \frac{1}{g_Y^{n-1}(y^{n-1})} \int_{y_n \in B} g_Y^n(y^n) d\eta(y_n) \leq z_{n-1} , \end{aligned}$$

where  $\{Z_n\}$  is a super-martingale if and only if  $\{-Z_n\}$  is a sub-martingale. From  $z_n \geq 0$ ,  $E[Z_n] = \int g^n(y^n) d\eta^n(y^n) \leq 1$ , and Doob's martingale convergence theorem below, we see  $\lim_{n \rightarrow \infty} \frac{1}{z_n}$  exists and finite, so that with probability one as  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{z_n} > 0, \quad \lim_{n \rightarrow \infty} \log \frac{1}{z_n} > -\infty, \quad \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{z_n} \geq 0 .$$

**Lemma 3 (Theorem 35.5 [1])** Let  $\{X_i\}$  be a sub-martingale. If  $K := \sup_n E[|X_n|] < \infty$ , then  $X_n \rightarrow X$  with probability one, where  $X$  is a random variable satisfying  $E[|X|] \leq 1$ .

Hence

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{d\mu_Y^n}{d\nu_Y^n}(y^n) \geq 0 .$$

On the other hand, for  $k = 1, 2, \dots$ ,  $\frac{d\nu_Y^n}{d\eta^n}(y^n) \geq w_k \frac{d\nu_k^n}{d\eta^n}(y^n)$ , so that

$$\begin{aligned} & \frac{1}{n} \log \frac{d\mu_Y^n}{d\nu_Y^n}(y^n) \\ & \leq -\frac{1}{n} \log w_k + \frac{1}{n} \log \frac{d\mu_k^n}{d\nu_k^n}(y^n) + \frac{1}{n} \log \frac{d\mu_Y^n}{d\mu_k^n}(y^n) \end{aligned} \quad (6)$$

with probability one as  $n \rightarrow \infty$ . Note that for each  $k = 1, 2, \dots$ , since  $B_k$  is a finite set, there exists a universal Bayesian measure  $Q_k$  associated with  $B_k$ , so that

$$\frac{1}{n} \log \frac{d\mu_k^n}{d\nu_k^n}(y^n) = \frac{1}{n} \log \frac{P_k^n(y^n)}{Q_k^n(y^n)} \rightarrow 0$$

with probability one as  $n \rightarrow \infty$ . On the other hand, from the law of large numbers,

$$\begin{aligned} & \frac{1}{n} \log \frac{d\mu_Y^n}{d\mu_k^n}(y^n) = \frac{1}{n} \sum_{i=1}^n \log \frac{d\mu_Y}{d\mu_k}(y_i) \\ & \rightarrow E[\log \frac{d\mu_Y}{d\mu_k}] = D(\mu_Y || \mu_k) . \end{aligned}$$

with probability one as  $n \rightarrow \infty$ . However, (6) should hold even for large  $k$ . From the assumption  $D(\mu_Y || \mu_k) \rightarrow 0$  as  $k \rightarrow \infty$ , we require

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{d\mu_Y^n}{d\nu_Y^n}(y^n) \leq 0 .$$

This completes the proof.

## 4 Universal Histogram Sequence

Theorem 1 assumes a specific  $\{B_k\}$ . Given  $\{B_k\}$ , for  $\mu$  such that  $D(\mu_k || \eta)$  does not converge to zero as  $k \rightarrow \infty$ , Eq. (5) does not hold in general. So, it is pleasing if we could find  $\{B_k\}$  such that for any  $\mu$ ,  $D(\mu_k || \eta) \rightarrow 0$  as  $k \rightarrow \infty$ . Then, we can remove Ryabko's condition.

To this end, we choose any  $\mu, \sigma \in \mathbb{R}$ , where  $\sigma$  should be positive, and generate the following sequence:

$$C_0 = \{(-\infty, \infty)\}$$

$$C_1 = \{(-\infty, \mu], (\mu, \infty)\}$$

Given

$$C_k = \{(-\infty, c_{k,1}], (c_{k,1}, c_{k,2}], \dots, (c_{k,2^k-2}, c_{k,2^k-1}], (c_{k,2^k-1}, \infty)\}$$

for  $k \geq 1$ , we define

$$\begin{aligned} C_{k+1} = & \{(-\infty, c_{k+1,1}], (c_{k+1,1}, c_{k+1,2}], \\ & \dots, (c_{k+1,2^{k+1}-2}, c_{k+1,2^{k+1}-1}], (c_{k+1,2^{k+1}-1}, \infty)\} \end{aligned}$$

by

$$c_{k+1,1} = \mu - k\sigma, \quad c_{k+1,2^{k+1}-1} = \mu + k\sigma$$

$$c_{k+1,2j} = c_{k,j}, \quad j = 1, \dots, 2^k - 1$$

$$c_{k+1,2j+1} = \frac{c_{k,j} + c_{k,j+1}}{2}, \quad j = 1, \dots, 2^k - 2$$

Therefore,  $C_k$  contains  $2^k$  elements. In this way, given the values of  $\mu, \sigma$ , we obtain the sequence  $\{C_k\}_{k=0}^\infty$ .

Let  $B$  be the set in which random variable  $Y$  takes values, and define

$$B_k^* := \{B \cap c \mid c \in C_k\} \setminus \{\phi\}.$$

**Example 6** Let  $B$  be the entire real  $\mathbb{R}$  ( $\{B_k^*\} = \{C_k\}$ ). We assume that  $\eta$  is the Lebesgue measure  $\lambda$ , and that  $\mu_Y \ll \lambda$  for  $\mu_Y(D) = P(Y \in D)$  for  $D \in \mathcal{B}$ , which means from Lemma 2 that a density function  $f_Y$  exists. Thus, for each  $y \in B = \mathbb{R}$ , there exist a unique sequence  $\{(a_k, b_k]\}_{k=1}^\infty$  such that  $-\infty < a_k \leq y \leq b_k < \infty$ ,  $k = 1, 2, \dots$ , where  $a_k, b_k = \pm\infty$  are allowed, so that the ratio

$$\frac{\mu_Y((a_k, b_k])}{\lambda((a_k, b_k])} = \frac{F_Y(b_k) - F_Y(a_k)}{b_k - a_k}$$

converges to  $f(y)$  as  $k \rightarrow \infty$ . Thus,  $D(f||f_k) \rightarrow 0$  as  $k \rightarrow \infty$  for any  $f$ , where  $F_Y$  is the distribution function of  $Y$ .

**Example 7** The sequence  $\{B_k\}$  in Example 5 is obtained by  $\mu = 1$  and  $\sigma = 1$  in the histogram sequence  $\{B_k^*\}$ . Then, for each  $y \in B = \mathbb{N} = \{1, 2, \dots\}$ , there exists  $K \in \mathbb{N}$  and a unique  $\{D_k\}_{k=1}^\infty$  such that  $y \in D_k \in B_k^*$ ,  $k = 1, 2, \dots$  and  $\{y\} = D_k \in B_k^*$  for  $k = K, K+1, \dots$ , so that

$$f_k(y) = \frac{\mu_Y(D_k)}{\eta(D_k)} \rightarrow f(y) = \frac{\mu_Y(\{y\})}{\eta(\{y\})}$$

for each  $y \in B$  and  $D(f||f_k) \rightarrow 0$  as  $k \rightarrow \infty$  for any  $f$ .

The choice of  $\mu, \sigma$  may be arbitrary, but we should take the prior knowledge into consideration in order to make the estimation correct even for small  $n$ .

**Theorem 2** If  $\mu \ll \eta$ , there exists a  $\nu \ll \eta$  such that  $\nu^n(B^n) \leq 1$  and with probability one as  $n \rightarrow \infty$ , (5) holds for any  $\mu$ .

Proof. It is sufficient to show  $D(f||f_k) \rightarrow 0$  as  $k \rightarrow \infty$  for the histogram sequence  $\{B_k^*\}$ . To this end, we consult the following lemma:

**Lemma 4 ([1], Problem 32.13)** Let  $\mu$  be the probability measure over  $\mathcal{B}$ ,  $\eta$  a  $\sigma$ -finite measure such that  $\mu \ll \eta$ . Then, with probability one,

$$\lim_{h \rightarrow 0} \frac{\mu((x-h, x+h])}{\eta((x-h, x+h])} = f(x),$$

where  $f$  is the density function of  $\mu$  w.r.t.  $\eta$ .

(For proof, see Appendix B.)

In our case, for each  $y \in B$ , there exists a unique sequence  $\{(a_k, b_k]\}_{k=1}^\infty$  such that  $y \in (a_k, b_k] \in B_k^*$ ,  $k = 1, 2, \dots$  and  $|b_k - a_k| \rightarrow 0$  ( $k \rightarrow \infty$ ), and obtain

$$\lim_{k \rightarrow \infty} \frac{\mu_Y((a_k, b_k])}{\eta((a_k, b_k])} = f(y)$$

with probability one. Hence,  $D(f||f_k) \rightarrow 0$  as  $k \rightarrow \infty$ . This completes the proof.

Hereafter, we refer  $\{B_k^*\}$  to the universal histogram sequence w.r.t.  $B$ .



## 5 When more than one variable exist

Analogous to the one variable case, we apply the notion of estimating Radon-Nikodym derivatives to the two random variables case.

Let  $\mu_X(D_X) := P(X \in D_X)$  and  $\mu_Y(D_Y) := P(Y \in D_Y)$  for  $D_X, D_Y \in \mathcal{B}$ , and  $\eta_X, \eta_Y$   $\sigma$ -finite measures such that  $\mu_X \ll \eta_X$  and  $\mu_Y \ll \eta_Y$ , respectively. Then, for<sup>3</sup>  $\mu_{XY}(D_X \times D_Y) = P(X \in D_X, Y \in D_Y)$ ,  $D_X, D_Y \in \mathcal{B}$ , we have  $\mu_{XY} \ll \eta_X \times \eta_Y$ , where  $\eta_X \times \eta_Y$  is the product measure of  $\eta_X, \eta_Y$ :  $\eta_X \times \eta_Y(D_X \times D_Y) = \eta_X(D_X)\eta_Y(D_Y)$ . Hence, from Lemma 2, there exists  $f_{XY}$  such that

$$\mu_{XY}(D_X \times D_Y) = \int_{x \in D_X} \int_{y \in D_Y} f_{XY}(x, y) d\eta_X(x) d\eta_Y(y) .$$

Let  $\{A_j\}, \{B_k\}$  be such that  $A_0 = \{A\}$  and  $B_0 = \{B\}$  and  $A_{j+1}, B_{k+1}$  are refinements of  $A_j, B_k$ , where  $A, B$  are the sets in which  $X, Y$  take values, respectively.

For each  $j, k = 1, 2, \dots$ ,  $s_j : A \rightarrow A$  and  $t_k : B \rightarrow B$  be such that  $x \in a \in A_j \implies s_j(x) = a$  and  $y \in a \in B_k \implies t_k(y) = b$ , respectively,

$$P_{jk}(a, b) := \int_{x \in a} \int_{y \in b} f_{XY}(x, y) d\eta_X(x) d\eta_Y(y)$$

for  $a \in A_j$  and  $b \in B_k$ , and

$$f_{jk}(x, y) := \frac{P_{jk}(s_j(x), t_k(y))}{\eta_X(s_j(x))\eta_Y(t_k(y))}$$

for  $x \in A$  and  $y \in B$ .

Since  $A_j \times B_k$  is a finite set, we can construct a universal Bayesian measure  $Q_{j,k}$  associated with  $A_j \times B_k$ .

Suppose we are given  $x^n = (x_1, \dots, x_n) \in A^n$  and  $y^n = (y_1, \dots, y_n) \in B^n$  such that  $(s_j(x_1), \dots, s_j(x_n)) = (a_1, \dots, a_n)$  and  $(t_k(y_1), \dots, t_k(y_n)) = (b_1, \dots, b_n)$  for  $j, k = 1, 2, \dots$ . Then, for each  $j, k = 1, 2, \dots$ , we estimate

$$\begin{aligned} f_{j,k}^n(x^n, y^n) &= f_{j,k}(x_1, y_1) \cdots f_{j,k}(x_n, y_n) \\ &= \frac{P_{j,k}(a_1, b_1) \cdots P_{j,k}(a_n, b_n)}{\eta_X(a_1) \cdots \eta_X(a_n) \eta_Y(b_1) \cdots \eta_Y(b_n)} \end{aligned}$$

by

$$g_{j,k}^n(x^n, y^n) = \frac{Q_{j,k}^n(a_1, b_1, \dots, a_n, b_n)}{\eta_X(a_1) \cdots \eta_X(a_n) \eta_Y(b_1) \cdots \eta_Y(b_n)} .$$

Then, we obtain

$$f_Y^n(x^n, y^n) = f_Y(x_1, y_1) \cdots f_Y(x_n, y_n)$$

and

$$g_Y^n(x^n, y^n) = \sum_{j,k} w_{j,k} g_{j,k}(x^n, y^n)$$

for some  $\{w_{j,k}\}$  such that  $w_{j,k} > 0$  and  $\sum_{j,k} w_{j,k} = 1$ .

As  $\{A_j \times B_k\}$ , we use  $\{(A_j \times B_k)^*\}$  defined by

$$(A_j \times B_k)^* := \{(A \cap c) \times (B \cap d) | c \in C_j, d \in C_k\} \setminus \{\phi\}$$

**Theorem 3** Suppose  $\mu_X \ll \eta_X$  and  $\mu_Y \ll \eta_Y$ , there exists  $\nu_{XY}^n \ll \eta_X^n \times \eta_Y^n$  such that  $\nu^n(A^n \times B^n) \leq 1$  and with probability one as  $n \rightarrow \infty$

$$\frac{1}{n} \log \frac{d\mu_{XY}^n}{d\nu_{XY}^n}(x^n, y^n) \rightarrow 0$$

for any  $\mu_{XY}$ .

---

<sup>3</sup>  $A \times B$  denotes the Cartesian product of sets  $A, B$ .

Proof: First, we notice for each  $x^{n-1} = (x_1, \dots, x_{n-1})$  and  $y^{n-1} = (y_1, \dots, y_{n-1})$

$$\int_{x_n \in A} \int_{y_n \in B} g_{j,k}^n(x^n, y^n) d\eta_X(x_n) d\eta_Y(y_n) \leq g_{j,k}^{n-1}(x^{n-1}, y^{n-1})$$

By weighting by  $\sum_{j,k} w_{j,k}[\cdot]$  for the both sides, we have

$$\int_{y_n} g_{XY}^n(x^n, y^n) d\eta_X(x_n) d\eta_Y(y_n) \leq g_{XY}^{n-1}(x^{n-1}, y^{n-1}) ,$$

so that  $\{Z_n\}$  with  $z_n := \frac{g_{XY}^n(x^n, y^n)}{f_{XY}^n(x^n, y^n)}$  is a super-martingale:

$$\begin{aligned} & E[Z|x^{n-1}, y^{n-1}] \\ &= \frac{g_{XY}^{n-1}(x^{n-1}, y^{n-1})}{f_{XY}^{n-1}(x^{n-1}, y^{n-1})} \cdot E\left[\frac{g_{XY}^n(x^n, y^n)}{g_{XY}^{n-1}(x^{n-1}, y^{n-1})} \cdot \frac{1}{f_{XY}(x_n, y_n)}\right] \\ &\leq z_{n-1} \cdot \frac{1}{g_{XY}^{n-1}(x^{n-1}, y^{n-1})} \\ &\quad \cdot \int_{x_n \in A} \int_{y_n \in B} g_{XY}^n(x^n, y^n) d\eta_X(x_n) d\eta_Y(y_n) \\ &\leq z_{n-1} . \end{aligned}$$

From  $z_n \geq 0$ ,  $E[Z_n] = \int g_{XY}^n(x^n, y^n) d\eta_X(x^n) d\eta_Y(y^n) \leq 1$ , and Doob's martingale convergence theorem, we see  $\lim_{n \rightarrow \infty} \frac{1}{z_n}$  exists and finite, so that with probability one as  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{z_n} > 0, \lim_{n \rightarrow \infty} \log \frac{1}{z_n} > -\infty, \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{z_n} \geq 0 .$$

Hence

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{d\mu_{XY}^n}{d\nu_{XY}^n}(x^n, y^n) \geq 0 .$$

On the other hand, for  $j, k = 1, 2, \dots$ ,

$$\frac{d\nu_{XY}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n) \geq w_{j,k} \frac{d\nu_{j,k}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n) ,$$

so that

$$\begin{aligned} & \frac{1}{n} \log \frac{d\mu_{XY}^n}{d\nu_{XY}^n}(x^n, y^n) \\ &\leq -\frac{1}{n} \log w_{j,k} + \frac{1}{n} \log \frac{d\mu_{j,k}^n}{d\nu_{j,k}^n}(x^n, y^n) \\ &\quad + \frac{1}{n} \log \frac{d\mu_{XY}^n}{d\mu_{j,k}^n}(x^n, y^n) \end{aligned}$$

with probability one as  $n \rightarrow \infty$ . Note that for each  $j, k = 1, 2, \dots$ , since  $A_j \times B_k$  is a finite set, there exists a universal Bayesian measure  $Q_{j,k}$  associated with  $A_j \times B_k$ , so that

$$\frac{1}{n} \log \frac{d\mu_{j,k}^n}{d\nu_{j,k}^n}(x^n, y^n) = \frac{1}{n} \log \frac{P_{j,k}^n(x^n, y^n)}{Q_{j,k}^n(x^n, y^n)} \rightarrow 0$$

with probability one as  $n \rightarrow \infty$ . On the other hand,

$$\begin{aligned} & \frac{1}{n} \log \frac{d\mu_{XY}^n}{d\mu_k^n}(x^n, y^n) = \frac{1}{n} \sum_{i=1}^n \log \frac{d\mu_{XY}}{d\mu_{j,k}}(x_i, y_i) \\ &\rightarrow E\left[\log \frac{d\mu_{XY}}{d\mu_{j,k}}\right] = D(\mu_{XY} || \mu_{j,k}) . \end{aligned}$$

So, it is sufficient to show  $D(\mu_{XY} || \mu_{j,k}) \rightarrow 0$  as  $j, k \rightarrow \infty$  for histogram sequence  $\{(A_j \times B_k)^*\}$ .

To this end, we consult the following lemma:

**Lemma 5** Let  $\mu_{XY}$  be the probability measure over  $\mathcal{B}^2$ ,  $\eta_X, \eta_Y$   $\sigma$ -finite measures such that  $\mu_{XY} \ll \eta_X$  and  $\mu_{XY} \ll \eta_Y$ . Then, with probability one,

$$\begin{aligned} & \lim_{h_x \rightarrow 0} \lim_{h_y \rightarrow 0} \frac{\mu_{XY}([x - h_x, x + h_x] \times [y - h_y, y + h_y])}{\eta_X((x - h_x, x + h_x])\eta_Y((y - h_y, y + h_y])} \\ &= f_{XY}(x, y), \end{aligned}$$

where  $f_{XY}$  is the density function of  $X, Y$ .

(For proof, see Appendix B.)

In our case, for each  $(x, y) \in A \times B$ , there exist a unique pair of sequences  $\{(a_j, b_j)\}_{j=1}^\infty$  and  $\{(c_k, d_k)\}_{k=1}^\infty$  such that  $x \in (a_j, b_j] \in A_j^*$ ,  $j = 1, 2, \dots$ ,  $y \in (c_k, d_k] \in B_k^*$ ,  $k = 1, 2, \dots$ ,  $|b_j - a_j| \rightarrow 0$  ( $j \rightarrow \infty$ ),  $|d_k - c_k| \rightarrow 0$  ( $k \rightarrow \infty$ ) and obtain

$$\lim_{j \rightarrow \infty} \lim_{k \rightarrow \infty} \frac{\mu_{XY}((a_j, b_j] \times (c_k, d_k])}{\eta_X((a_j, b_j])\eta_Y((c_k, d_k])} = f_{XY}(x, y)$$

with probability one. Hence,  $D(f || f_{j,k}) \rightarrow 0$  as  $j, k \rightarrow \infty$ . Thus,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{d\mu_{XY}^n}{d\nu_{XY}^n}(x^n, y^n) \leq 0.$$

This completes the proof.

**Example 8** Suppose  $\{A_j\}, \{B_k\}, \eta_X, \eta_Y$  are given by the universal histogram sequences w.r.t.  $A = [0, 1), B = \{1, 2, \dots\}$  and  $\eta_X = \lambda, \eta_Y = \eta$  in Examples 1 and 5. Then, we can construct  $\frac{d\nu_{XY}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n)$  from  $x^n \in A^n$  and  $y^n \in B^n$ .

It is straightforward to extend the result for the two variable case to the  $m(\geq 2)$  variables case.

## 6 Concluding Remarks

In this paper, we successfully construct a universal Bayesian measure for any random variables:

1. we extended Ryabko's measure so that the random variables may be either discrete or continuous or none of them;
2. constructed a universal histogram sequence that realizes universality for any source; and
3. constructed a universal Bayesian measure for more than one variables.

The results in this paper are rather theoretical but contain many applications such as

1. Bayesian network structure learning [7, 8],
2. a variant of the Chow-Liu algorithm learning a forest given examples [7, 9].

In fact, in any database, both discrete and continuous fields are present. Then, we need to find dependency among those attributes. However, the existing results only dealt with either only discrete data or only continuous data. This paper deals with the most general and realistic cases.

For contributions to statistics, constructing such a universal Bayesian measure means establishing a general form of Bayesian Information Criteria (BIC). Suppose we have a countable number of models  $m = 1, 2, \dots$  each of which expresses a relation among random variables. If we construct a universal Bayesian measure  $q(x^n | m)$  w.r.t. model  $m$  given data  $x^n$ , then we can select  $m$  such that  $-\log p(m) - \log q(x^n | m)$  is minimized, where  $p(m)$  is the prior probability of model  $m$ . In fact, the measure applies to all the cases that BIC/MDL applied thus far.

## Appendix A: Proof of existence of $Q$ satisfying (2) and (3) for finite set $A$

Although the proposition is standard [6], we give a proof for selfcontainedness.

Let  $c[x]$  be the frequency of  $x \in A$  in  $x^n = (x_1, \dots, x_n) \in A^n$ . Then, we see

$$Q_X^n(x^n) := \frac{\Gamma(\sum_{x \in A} a[x]) \prod_{x \in A} (c[x] + a[x])}{\Gamma(\sum_{x' \in A} (c[x'] + a[x'])) \prod_{x \in A} \Gamma(a[x])}$$

with  $a[x] = \frac{1}{2}$ ,  $x \in A$ , satisfies (2), where  $m = |A|$  and  $\Gamma$  is the Gamma function:  $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ . In fact,  $Q_X^0(x^0) = 1$ ; and if (2) is assumed for  $n \geq 0$ , then for  $x_{n+1} \in A$  and  $x^{n+1} = (x_1, \dots, x_{n+1}) \in A^{n+1}$ ,

$$\begin{aligned} & \sum_{x^{n+1} \in A^{n+1}} Q^{n+1}(x^{n+1}) \\ &= \sum_{x^{n+1} \in A^{n+1}} Q_X^n(x^n) \cdot \frac{c[x_{n+1}] + \frac{1}{2}}{n + \frac{m}{2}} \\ &= \sum_{x^n \in A^n} Q_X^n(x^n) \sum_{x_{n+1} \in A} \frac{c[x_{n+1}] + \frac{1}{2}}{n + \frac{m}{2}} \leq 1, \end{aligned}$$

where  $\Gamma(z+1) = z\Gamma(z)$  has been applied for  $z > 0$ . Thus, we obtain (2).

From Stirling's formula<sup>4</sup>:

$$\log \Gamma(z) = (z - \frac{1}{2}) \log z - z - \frac{1}{2} \log(2\pi) + o(1),$$

we have

$$\begin{aligned} & -\log Q_X^n(x^n) \\ &= (n + \frac{m-1}{2}) \log(n + \frac{m}{2}) - (n + \frac{m}{2}) \\ & \quad - \sum_{x \in A} c[x] \log(c[x] + \frac{1}{2}) + \sum_{x \in A} (c[x] + \frac{1}{2}) \\ & \quad + \frac{m-1}{2} \log(2\pi) + o(1) \\ &= - \sum_{x \in A} c[x] \log \frac{c[x] + \frac{1}{2}}{n + \frac{m}{2}} + \frac{m-1}{2} \log(n + \frac{m}{2}) + O(1) \end{aligned}$$

From the law of large numbers,  $c[x]/n$  converges to the probability  $P(x)$  of  $x \in A$  with probability one as  $n \rightarrow \infty$  for independent source  $P_X^n(x^n) = \prod_{i=1}^n P(x_i)$  for  $x^n = (x_1, \dots, x_n) \in A^n$ , so that  $-\log Q_X^n(x^n)$  converges to its entropy  $H(P) := \sum_{x \in A} -P(x) \log P(x)$ . On the other hand, from the law of large numbers, with probability one as  $n \rightarrow \infty$

$$\begin{aligned} & -\frac{1}{n} \log P_X^n(x^n) = \frac{1}{n} \sum_{i=1}^n \{-\log P(x_i)\} \\ & \rightarrow E[-\log P(X)] = H(P) \end{aligned}$$

(Shannon-McMillian-Breiman [2]). Thus, we obtain (3), and this completes the proof.

## Appendix B: Proof of Lemmas 4 and 5

For the one variable case, let  $y \in \mathbb{R}$ . Since  $\mu_Y \ll \eta$  for  $\mu_Y(D) = P(Y \in D)$ ,  $D \in \mathcal{B}$ , there exists  $f_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that for  $h > 0$ ,

$$\mu_Y([y-h, y+h]) = \int_{y-h}^{y+h} f_Y(u) d\eta(u)$$

---

<sup>4</sup>The natural logarithm is assumed.

Let  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  be such that

$$\varphi_Y(t) := \inf\{z \in \mathbb{R} | t \leq \eta^*([y, z])\},$$

where

$$\eta^*([y, z]) := \begin{cases} \eta([y, z]), & y \leq z \\ -\eta([z, y]), & y > z \end{cases}.$$

Then, the integral is expressed by

$$\int_{-\eta[y-h, y]}^{\eta[y, y+h]} f_Y(\varphi_Y(t)) dt$$

On the other hand, in general, for density function  $f$  and  $x \in \mathbb{R}$ , we have

$$\lim_{h \rightarrow 0} \frac{\int_{x-h}^{x+h} f(t) dt}{2h} = f(x)$$

Thus, as  $h \rightarrow 0$ , we have

$$\begin{aligned} \frac{\mu_Y([y-h, y+h])}{\eta([y-h, y+h])} &= \frac{\int_{-\eta[y-h, y]}^{\eta[y, y+h]} f_Y(\varphi_Y(t)) dt}{\eta([y, y+h]) - \{-\eta([y-h, y])\}} \\ \rightarrow f_Y \circ \varphi(0) &= f_Y(y) \end{aligned}$$

For the two variable case, let  $x, y \in \mathbb{R}$ . Since  $\mu_{XY} \ll \eta_X, \eta_Y$  for  $\mu_{XY}(D) = P(X \in D_X, Y \in D_Y)$ ,  $D_X, D_Y \in \mathcal{B}$ , there exists  $f_{XY} : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that for  $h_x, h_y > 0$

$$\begin{aligned} \mu_{XY}([x-h_x, x+h_x] \times [y-h_y, y+h_y]) \\ = \int_{x-h_x}^{x+h_x} \int_{y-h_y}^{y+h_y} f_{XY}(u, v) d\eta_X(u) d\eta_Y(v) \end{aligned}$$

Let  $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be such that

$$\begin{aligned} \varphi_Y(s, t) \\ := (\inf\{w \in \mathbb{R} | s \leq \eta^*([x, w])\}, \inf\{z \in \mathbb{R} | t \leq \eta^*([y, z])\}) \end{aligned}$$

Then, the integral is expressed by

$$\int_{-\eta_X[x-h_x, x]}^{\eta_X[x, x+h_x]} \int_{-\eta_Y[y-h_y, y]}^{\eta_Y[y, y+h_y]} f_{XY}(\varphi_{XY}(s, t)) ds dt$$

On the other hand, in general, for density function  $f$  and  $x, y \in \mathbb{R}$ , we have

$$\lim_{h_x \rightarrow 0} \lim_{h_y \rightarrow 0} \frac{\int_{x-h_x}^{x+h_x} \int_{y-h_y}^{y+h_y} f(s, t) ds dt}{4h_x h_y} = f(x, y)$$

Thus, as  $h_x, h_y \rightarrow 0$ , we have

$$\begin{aligned} \frac{\mu_Y([x-h_x, x+h_x] \times [y-h_y, y+h_y])}{\eta_X([x-h_x, x+h_x])\eta_Y([y-h_y, y+h_y])} \\ = \left\{ \int_{-\eta_X[x-h_x, x]}^{\eta_X[x, x+h_x]} \int_{-\eta_Y[y-h_y, y]}^{\eta_Y[y, y+h_y]} f_{XY}(\varphi_{XY}(s, t)) ds dt \right\} \\ / \{[\eta_X([x, x+h_x]) - \{-\eta_X([x-h_x, x])\}] \\ \cdot [\eta_Y([y, y+h_y]) - \{-\eta_Y([y-h_y, y])\}]\} \\ \rightarrow f_{XY} \circ \varphi(0, 0) = f_{XY}(x, y) \end{aligned}$$

This completes the proof.

## Acknowledgment

The authors would like to thank Prof. Boris Ryabko for suggesting me to develop further applications of his exciting theory.

## References

- [1] P. Billingsley. *Probability & Measure* (1995): (3rd ed.). New York : Wiley.
- [2] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (1995): (2nd ed.). New York : Wiley.
- [3] R.E. Krichevsky and V.K. Trofimov, “The Performance of Universal Encoding”, *IEEE Trans. Inform. Theory* 27(2): 199-207 (1981).
- [4] J.Rissanen, “Modeling by shortest data description”. *Automatica* 14: 465-471 (1978).
- [5] B. Ryabko, “Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series.” *IEEE Trans. on Inform. Theory*, 55(9):4309-4315 (2009).
- [6] B. Ryabko, “Prediction of random sequences and universal coding”, *Problems Inform. Transmission* 24 (1988), no. 2, 87–96. Russian: *Problemy Peredachi Informatsii* 24 (1988), no. 2,3–14
- [7] J. Suzuki, “A Construction of Bayesian Networks from Databases on an MDL Principle”, *The Ninth Conference on Uncertainty in Artificial Intelligence*, Washington D. C., pages 266-273, 7 (1993).
- [8] J. Suzuki, “The Universal Measure for General Sources and its Application to MDL/Bayesian Criteria”, *Data Compression Conference* 2011, Snowbird, Utah (2011).
- [9] J. Suzuki, “The Bayesian Chow-Liu Algorithmsh, pages 315-322, Proceedings of the 6th workshop on Probabilistic Graphical Models, Granada, Spain. (2012)